Prompt-to-Prompt Image Editing with Cross Attention Control

Year: 2022 | Citations: 2247 | Authors: Amir Hertz, Ron Mokady, J. Tenenbaum, Kfir Aberman, Y. Pritch

Abstract

Recent large-scale text-driven synthesis models have attracted much attention thanks to their remarkable capabilities of generating highly diverse images that follow given text prompts. Such text-based synthesis methods are particularly appealing to humans who are used to verbally describe their intent. Therefore, it is only natural to extend the text-driven image synthesis to text-driven image editing. Editing is challenging for these generative models, since an innate property of an editing technique is to preserve most of the original image, while in the text-based models, even a small modification of the text prompt often leads to a completely different outcome. State-of-the-art methods mitigate this by requiring the users to provide a spatial mask to localize the edit, hence, ignoring the original structure and content within the masked region. In this paper, we pursue an intuitive prompt-to-prompt editing framework, where the edits are controlled by text only. To this end, we analyze a text-conditioned model in depth and observe that the cross-attention layers are the key to controlling the relation between the spatial layout of the image to each word in the prompt. With this observation, we present several applications which monitor the image synthesis by editing the textual prompt only. This includes localized editing by replacing a word, global editing by adding a specification, and even delicately controlling the extent to which a word is reflected in the image. We present our results over diverse images and prompts, demonstrating high-quality synthesis and fidelity to the edited prompts.